Synthetic Data: The Secret Ingredient in **Better Language Models**

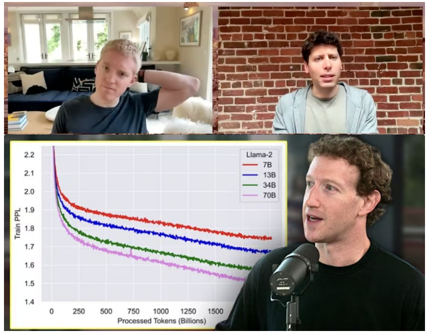
FOSDEM 2025: Low-Level AI DevRoom

Cedric Clyburn
Senior Developer Advocate
@cedricclyburn

Carol Chen
AI Community Architect

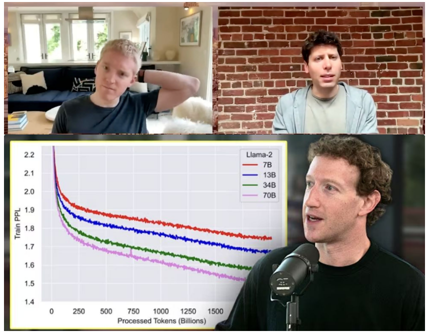
We’ve got **a lot** to cover today!

We’ve got **a lot** to cover today!



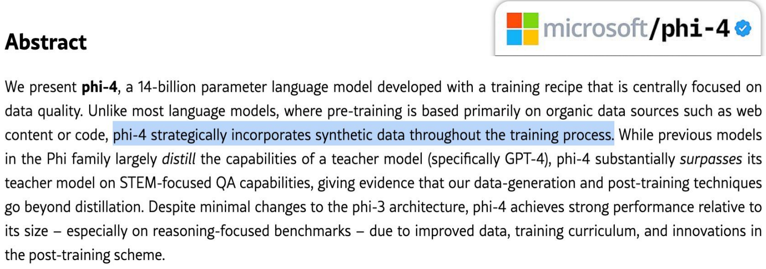
Why are the Zuck and Altman talking about synthetic data?

We’ve got **a lot** to cover today!



Why are the Zuck and Altman talking about synthetic data?

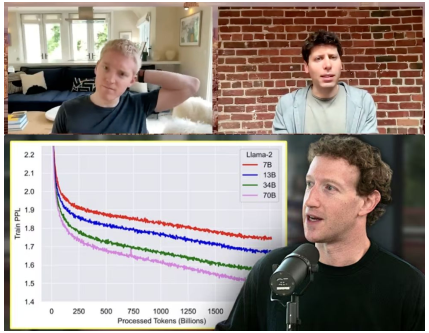
The newest foundation models use synthetic data!?



Abstract

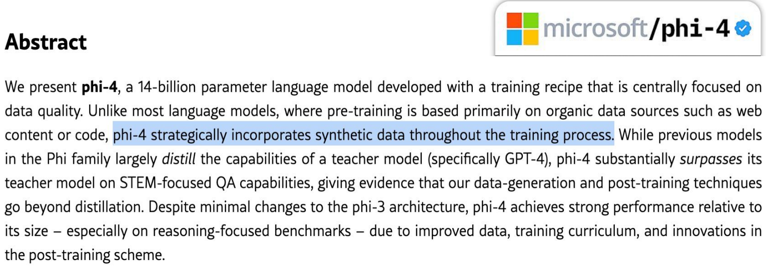
We present **phi-4**, a 14-billion parameter language model developed with a training recipe that is centrally focused on data quality. Unlike most language models, where pre-training is based primarily on organic data sources such as web content or code, phi-4 strategically incorporates synthetic data throughout the training process. While previous models in the Phi family largely *distill* the capabilities of a teacher model (specifically GPT-4), phi-4 substantially *surpasses* its teacher model on STEM-focused QA capabilities, giving evidence that our data-generation and post-training techniques go beyond distillation. Despite minimal changes to the phi-3 architecture, phi-4 achieves strong performance relative to its size – especially on reasoning-focused benchmarks – due to improved data, training curriculum, and innovations in the post-training scheme.

We've got **a lot** to cover today!



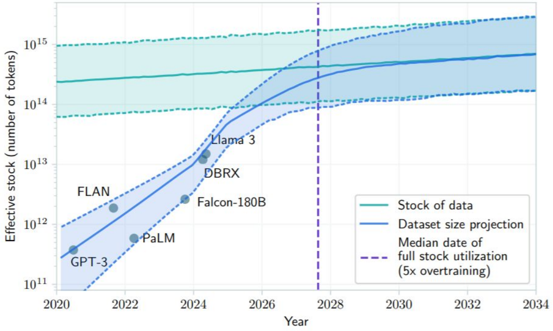
Why are the Zuck and Altman talking about synthetic data?

The newest foundation models use synthetic data!?



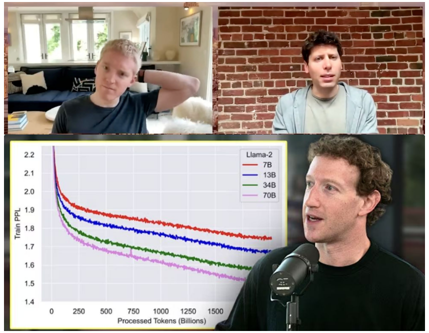
Abstract

We present **phi-4**, a 14-billion parameter language model developed with a training recipe that is centrally focused on data quality. Unlike most language models, where pre-training is based primarily on organic data sources such as web content or code, phi-4 strategically incorporates synthetic data throughout the training process. While previous models in the Phi family largely *distill* the capabilities of a teacher model (specifically GPT-4), phi-4 substantially *surpasses* its teacher model on STEM-focused QA capabilities, giving evidence that our data-generation and post-training techniques go beyond distillation. Despite minimal changes to the phi-3 architecture, phi-4 achieves strong performance relative to its size – especially on reasoning-focused benchmarks – due to improved data, training curriculum, and innovations in the post-training scheme.



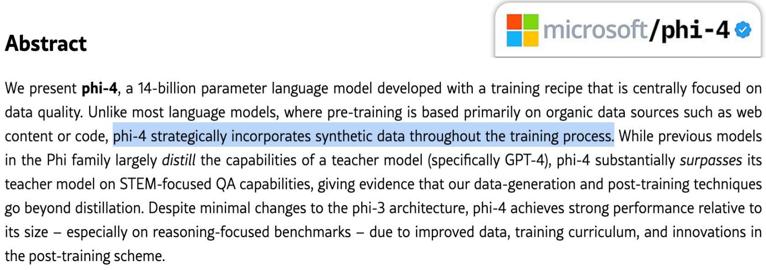
Will we soon run out of training data?!

We've got **a lot** to cover today!



Why are the Zuck and Altman talking about synthetic data?

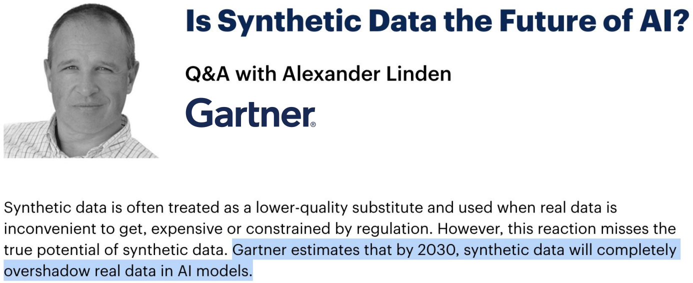
The newest foundation models use synthetic data!?



Abstract

We present **phi-4**, a 14-billion parameter language model developed with a training recipe that is centrally focused on data quality. Unlike most language models, where pre-training is based primarily on organic data sources such as web content or code, phi-4 strategically incorporates synthetic data throughout the training process. While previous models in the Phi family largely *distill* the capabilities of a teacher model (specifically GPT-4), phi-4 substantially *surpasses* its teacher model on STEM-focused QA capabilities, giving evidence that our data-generation and post-training techniques go beyond distillation. Despite minimal changes to the phi-3 architecture, phi-4 achieves strong performance relative to its size – especially on reasoning-focused benchmarks – due to improved data, training curriculum, and innovations in the post-training scheme.

There'll be more synthetic data than real data in models?

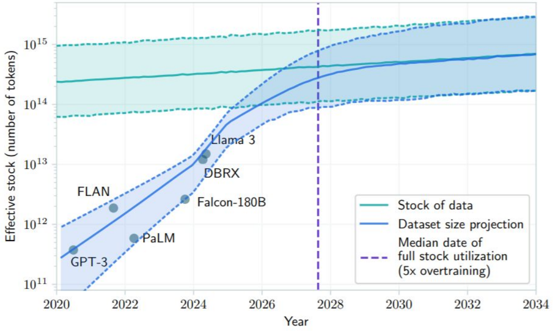


Is Synthetic Data the Future of AI?

Q&A with Alexander Linden

Gartner

Synthetic data is often treated as a lower-quality substitute and used when real data is inconvenient to get, expensive or constrained by regulation. However, this reaction misses the true potential of synthetic data. Gartner estimates that by 2030, synthetic data will completely overshadow real data in AI models.



Will we soon run out of training data?!

Agenda

Today's Schedule

Slides

red.ht/synthetic-slides

- Challenges with AI in Production
- The Role of **Synthetic Data**
- Where it fits in the LLM Pipeline
- **Demo #1:** Language Classification!
- **Demo #2:** Domain Specific SLM
- Resources, Links, Q&A!

& a quick introduction to your speakers!

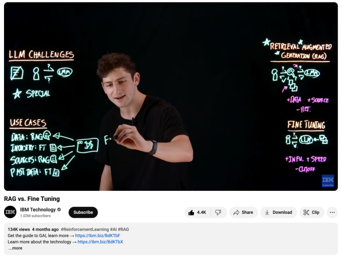
Cedric Clyburn

Senior Developer

Advocate

KCD
GROWING CLOUD NATIVE TOGETHER
ORGANIZER
2024

AI.dev
SPEAKER
2023

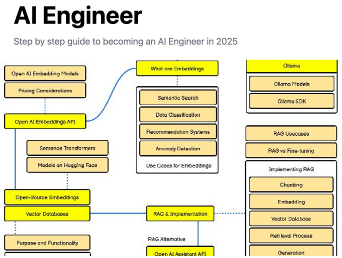


RAG vs. Fine Tuning

IBM Technology
1.02M subscribers
Subscribe

4.4K Share Download Clip

134K views 4 months ago #ReinforcementLearning #AI #RAG
Get the guide to GAI, learn more → https://ibm.biz/BdKTbF
Learn more about the technology → https://ibm.biz/BdKTbX
...more



AI Engineer

Step by step guide to becoming an AI Engineer in 2025

Open AI Embedding Models
Pricing Considerations
Open AI Embeddings API
Sentence Transformers
Models on Hugging Face
Open-Source Embeddings
Vector Databases
Purpose and Functionality
What are Embeddings
Semantic Search
Data Classification
Recommendation Systems
Anomaly Detection
Use Cases for Embeddings
RAG & Implementation
RAG Alternative
Open AI Assistant API
Ollama
Ollama Models
Ollama SDK
RAG Usecases
RAG vs Fine-tuning
Implementing RAG
Chunking
Embedding
Vector Database
Retrieval Process
Generation

roadmap.sh/ai-engineer

Carol Chen

AI Community

Architect

MeeGo

InstructLab

SAILFISH OS

ManageIQ

@cybette
(@mastodon.org.uk, github,
:matrix.org, bsky.social)

What are common challenges for AI adoption?

Our analysis, from working with Fortune 500 and small organizations

Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise use cases.

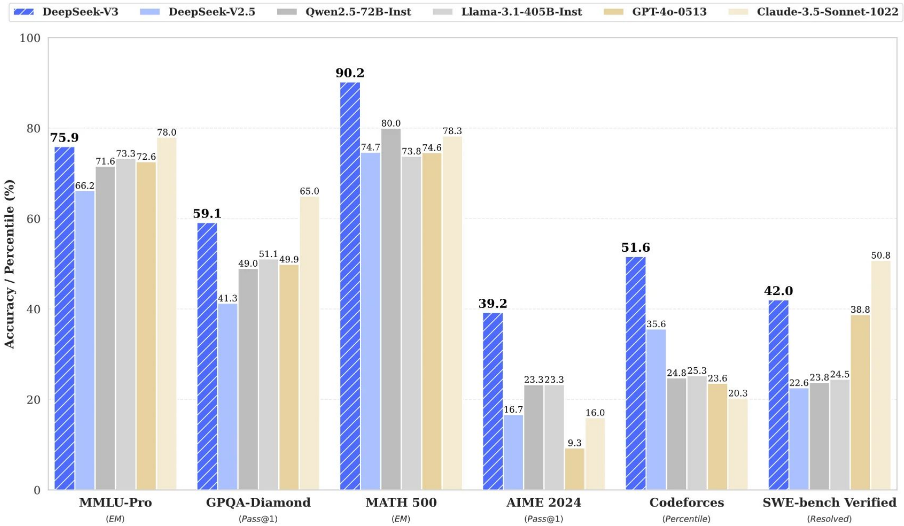
Complexity

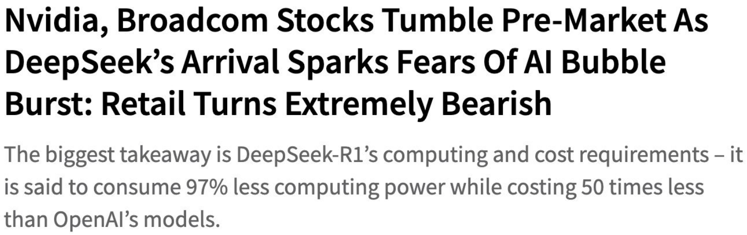
Tuning models with private data for enterprise use cases is too complex for non-data scientists.

Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

Well, speaking about cost...



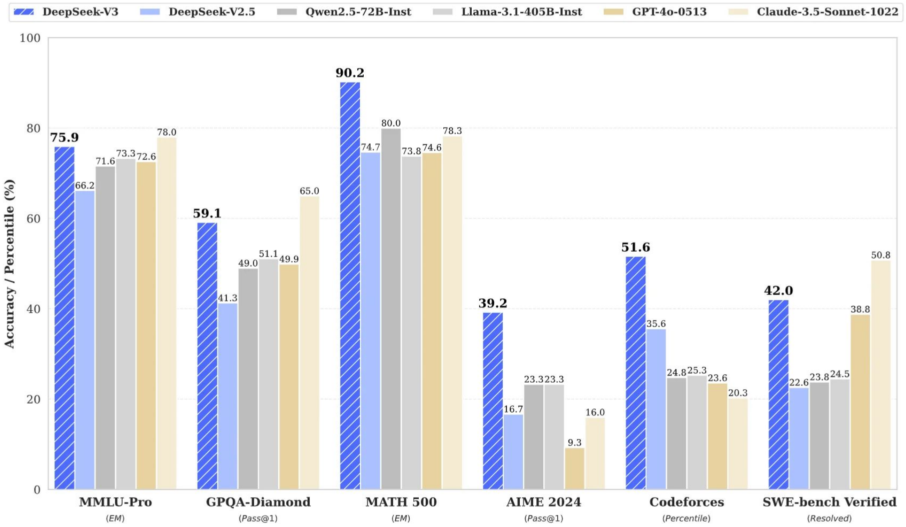


Nvidia, Broadcom Stocks Tumble Pre-Market As DeepSeek's Arrival Sparks Fears Of AI Bubble Burst: Retail Turns Extremely Bearish

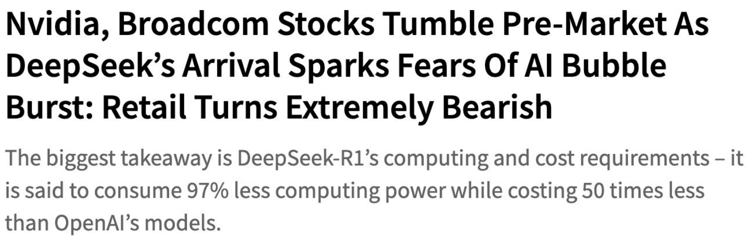
The biggest takeaway is DeepSeek-R1's computing and cost requirements – it is said to consume 97% less computing power while costing 50 times less than OpenAI's models.

deepseek

Well, speaking about cost...



Outperforms closed source on many benchmarks

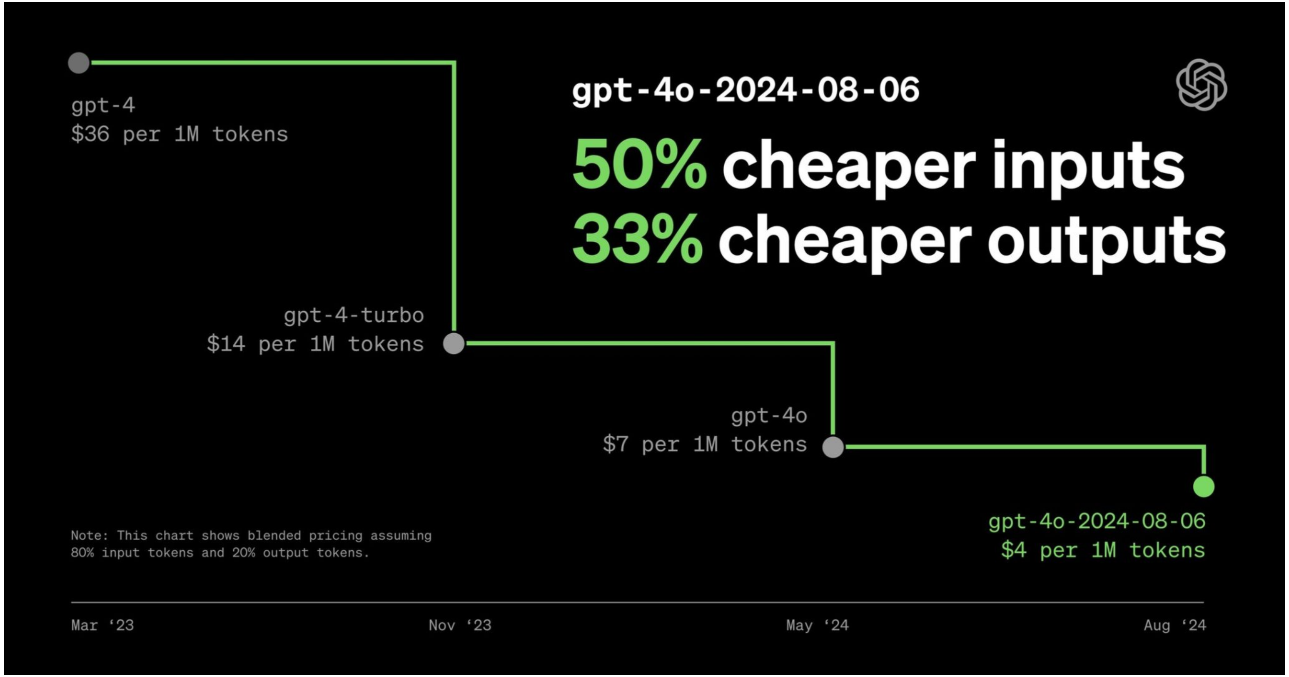


Nvidia, Broadcom Stocks Tumble Pre-Market As DeepSeek's Arrival Sparks Fears Of AI Bubble Burst: Retail Turns Extremely Bearish

The biggest takeaway is DeepSeek-R1's computing and cost requirements – it is said to consume 97% less computing power while costing 50 times less than OpenAI's models.

deepseek

…and this trend isn't new!



gpt-4o-2024-08-06

50% cheaper inputs
33% cheaper outputs

Note: This chart shows blended pricing assuming 80% input tokens and 20% output tokens.

Models are becoming more cost-effective and efficient thanks to competition!

What are common challenges for AI adoption?

Our analysis, from working with Fortune 500 and small organizations

Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise use cases.

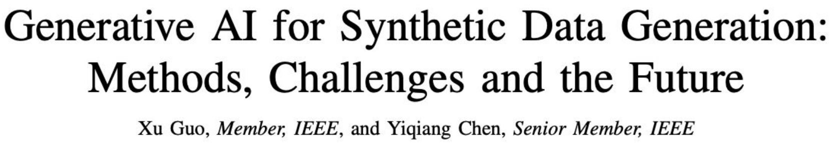
Complexity

Tuning models with private data for enterprise use cases is too complex for non-data scientists.

Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

AI-Enabled applications **need** to incorporate additional knowledge



Generative AI for Synthetic Data Generation: Methods, Challenges and the Future

Xu Guo, *Member, IEEE*, and Yiqiang Chen, *Senior Member, IEEE*

The necessity for synthetic data arises from the ***inherent limitations of general-purpose Large Language Models (LLMs) in specialized and private domains****, despite their significant achievements across various benchmarks; for instance, ClinicalBERT [17], adapted from BERT through pre-training on clinical texts, demonstrates* ***superior performance*** *in predicting hospital readmissions compared to the original BERT [18], which was trained on Wikipedia and BookCorpus [19] text data,* ***highlighting a crucial challenge: specialized domains often rely on domain-specific data that is not readily available or open to the public****, thereby underscoring the importance of synthetic data in bridging these gaps.*

AI-Enabled applications **need** to incorporate additional knowledge

Generative AI for Synthetic Data Generation: Methods, Challenges and the Future

Xu Guo, *Member, IEEE*, and Yiqiang Chen, *Senior Member, IEEE*

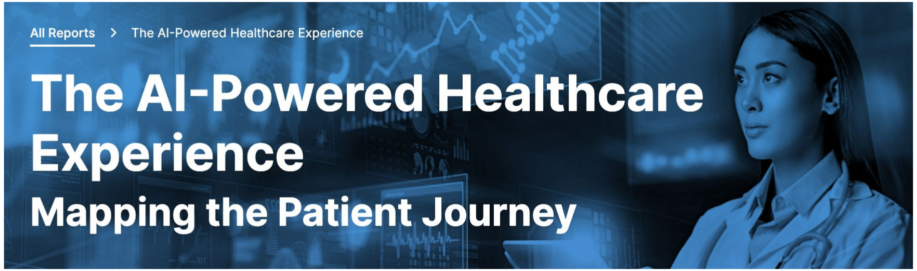
The necessity for synthetic data arises from the ***inherent limitations of general-purpose Large Language Models (LLMs) in specialized and private domains****, despite their significant achievements across various benchmarks; for instance, ClinicalBERT [17], adapted from BERT through pre-training on clinical texts, demonstrates* ***superior performance*** *in predicting hospital readmissions compared to the original BERT [18], which was trained on Wikipedia and BookCorpus [19] text data,* ***highlighting a crucial challenge: specialized domains often rely on domain-specific data that is not readily available or open to the public****, thereby underscoring the importance of synthetic data in bridging these gaps.*

One of the biggest limitations of LLMs are their general-ness, because specialized use cases require domain-specific data

AI-Enabled applications **need** to incorporate additional knowledge

svb Silicon Valley Bank

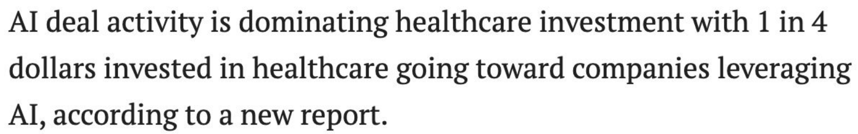
A Division of First Citizens Bank



All Reports > The AI-Powered Healthcare Experience

The AI-Powered Healthcare Experience

Mapping the Patient Journey

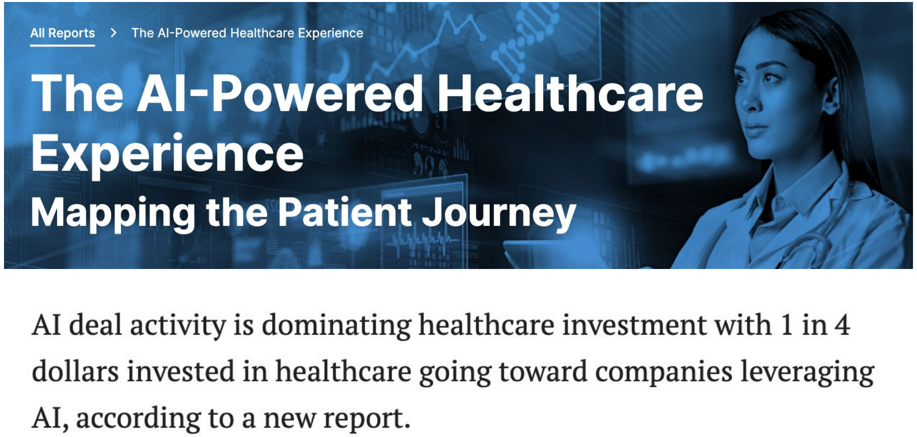


AI deal activity is dominating healthcare investment with 1 in 4 dollars invested in healthcare going toward companies leveraging AI, according to a new report.

AI-Enabled applications **need** to incorporate additional knowledge

svb Silicon Valley Bank

A Division of First Citizens Bank



All Reports > The AI-Powered Healthcare Experience

The AI-Powered Healthcare Experience

Mapping the Patient Journey

AI deal activity is dominating healthcare investment with 1 in 4 dollars invested in healthcare going toward companies leveraging AI, according to a new report.

But when it comes to regulated or sensitive industries, there isn't as much data accessible because of privacy

What are common challenges for AI adoption?

Our analysis, from working with Fortune 500 and small organizations

Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise use cases.

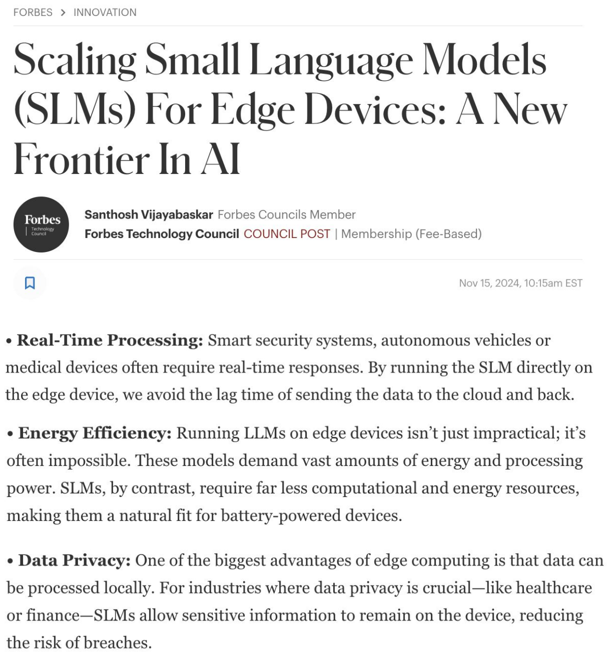
Complexity

Tuning models with private data for enterprise use cases is too complex for non-data scientists.

Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

Models can't be constrained to a single cloud service, and SLM's can run on-device...



FORBES > INNOVATION

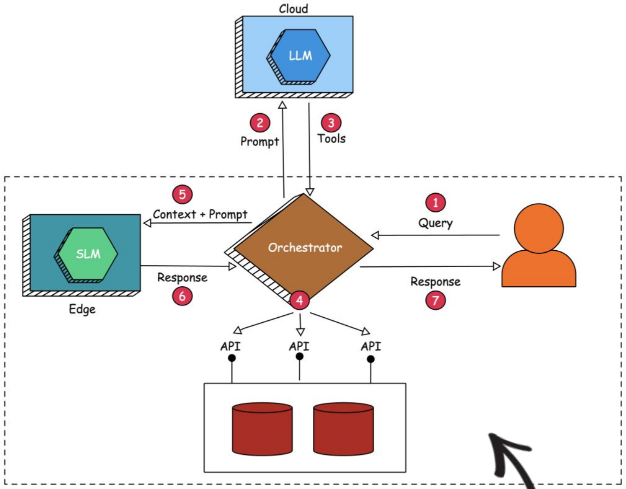
Scaling Small Language Models (SLMs) For Edge Devices: A New Frontier In AI

Santhosh Vijayabaskar Forbes Councils Member
Forbes Technology Council COUNCIL POST | Membership (Fee-Based)

Nov 15, 2024, 10:15am EST

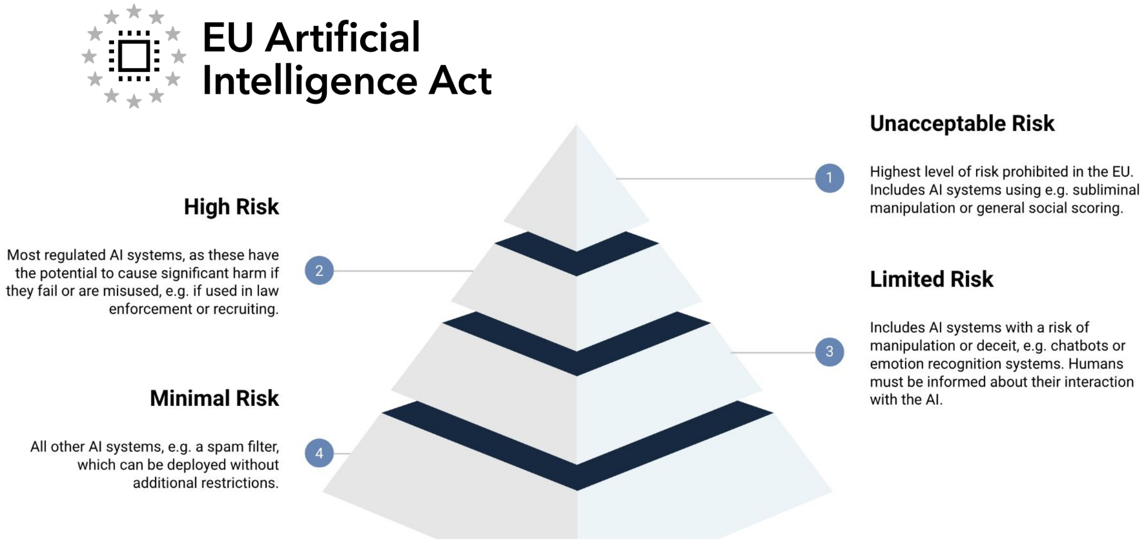
- **Real-Time Processing:** Smart security systems, autonomous vehicles or medical devices often require real-time responses. By running the SLM directly on the edge device, we avoid the lag time of sending the data to the cloud and back.
- **Energy Efficiency:** Running LLMs on edge devices isn't just impractical; it's often impossible. These models demand vast amounts of energy and processing power. SLMs, by contrast, require far less computational and energy resources, making them a natural fit for battery-powered devices.
- **Data Privacy:** One of the biggest advantages of edge computing is that data can be processed locally. For industries where data privacy is crucial—like healthcare or finance—SLMs allow sensitive information to remain on the device, reducing the risk of breaches.

There are plenty of situations (ex. driverless cars, Apple Intelligence) where models need to run on edge



SLM's can handle plenty of use-cases, and if not, LLM's can be supplemented

...which can be helpful when dealing with these tricky regulations!



EU Artificial Intelligence Act

Unacceptable Risk

1 Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

2 Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

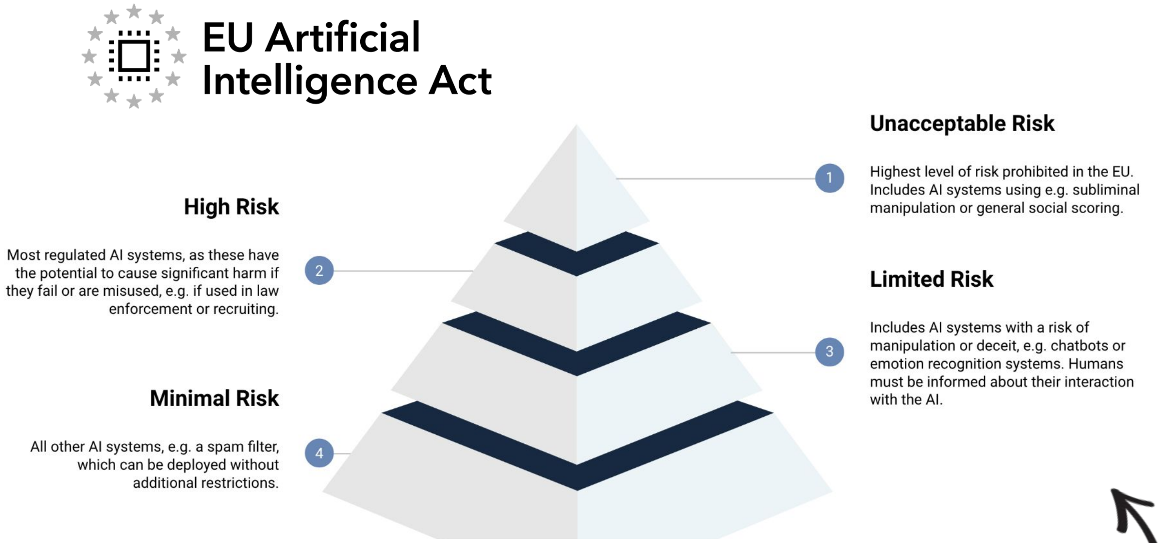
3 Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

4 All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

GDPR

...which can be helpful when dealing with these tricky regulations!



EU Artificial Intelligence Act

Unacceptable Risk

1 Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

2 Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

3 Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

4 All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

GDPR

Encourages data anonymization and synthetic data generation

Stringent requirements on data handling due to data privacy laws

What are common challenges for AI adoption?

Our analysis, from working with Fortune 500 and small organizations

Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise use cases.

Complexity

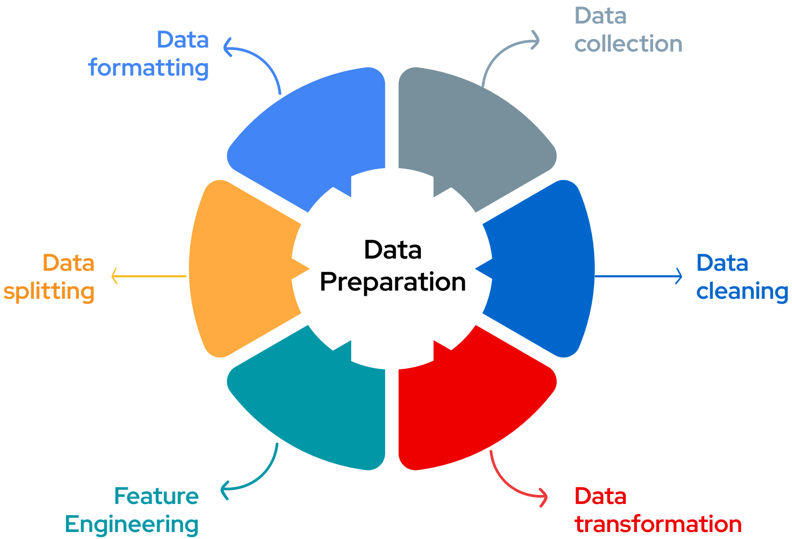
Tuning models with private data for enterprise use cases is too complex for non-data scientists.

Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

The process of data curation can be quite difficult

To develop a customized model requires a customized dataset



- Organizations can be challenged with having **enough data** to build an effective model
- **Synthetic data generation is becoming necessary** for accelerating model development, and overcoming data shortage.

What Is Synthetic Data?

Information that's been generated by a computer to augment or replace real data

Domain-Specific Data

Tailored to specialized industries and situations where real-world data is scarce.

Privacy-Preserving

Supplements real-world data without exposing confidential information.

Cost & Efficiency

Reduces the time, expense, and legal hurdles of collecting and labeling large datasets.

Data Quality

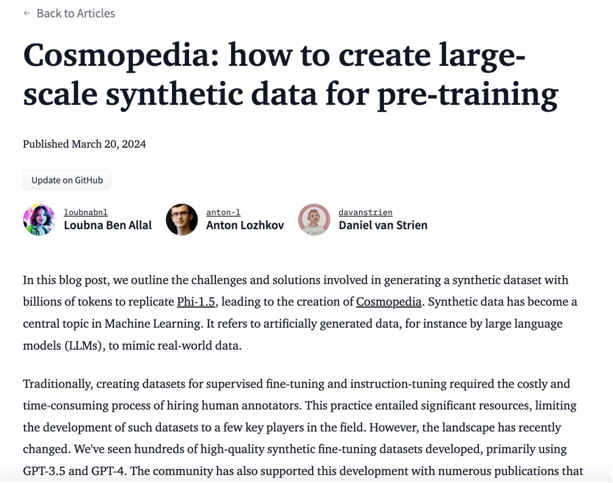
Perfectly annotated and high quality datasets, only limited by computational bottlenecks.

Where Synthetic Data is being used in the LLM Pipeline

You’ve likely already interacted with it!

Stage of Model Development	Use Case
Pre-Training for Foundation Models	Refining, classifying, and filtering the web to remove duplicated data, biases, and in general curate better-performing datasets (as well as address the model collapse challenge).

Synthetic representations of **pre-training data**



← Back to Articles

Cosmopedia: how to create large-scale synthetic data for pre-training

Published March 20, 2024

Update on GitHub

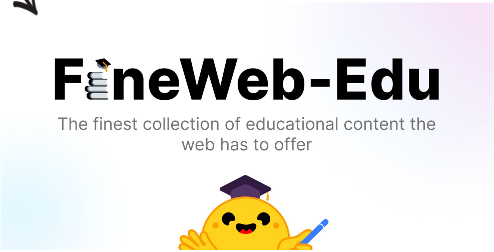
loubnabnl Loubna Ben Allal · anton-l Anton Lozhkov · davanstrien Daniel van Strien

In this blog post, we outline the challenges and solutions involved in generating a synthetic dataset with billions of tokens to replicate Phi-1.5, leading to the creation of Cosmopedia. Synthetic data has become a central topic in Machine Learning. It refers to artificially generated data, for instance by large language models (LLMs), to mimic real-world data.

Traditionally, creating datasets for supervised fine-tuning and instruction-tuning required the costly and time-consuming process of hiring human annotators. This practice entailed significant resources, limiting the development of such datasets to a few key players in the field. However, the landscape has recently changed. We've seen hundreds of high-quality synthetic fine-tuning datasets developed, primarily using GPT-3.5 and GPT-4. The community has also supported this development with numerous publications that

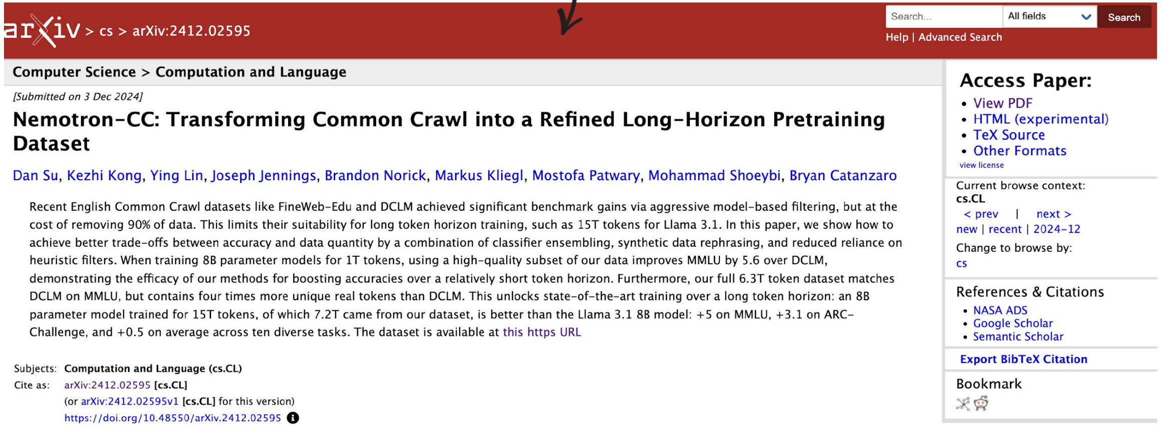
HuggingFace's experiment of prompting public webpages into a synthetic dataset

Annotating/filtering the web



The finest collection of educational content the web has to offer

Rewriting 1.9T tokens of data to remove low-quality data/diversify data



Computer Science > Computation and Language

[Submitted on 3 Dec 2024]

Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro

Recent English Common Crawl datasets like FineWeb-Edu and DCLM achieved significant benchmark gains via aggressive model-based filtering, but at the cost of removing 90% of data. This limits their suitability for long token horizon training, such as 15T tokens for Llama 3.1. In this paper, we show how to achieve better trade-offs between accuracy and data quantity by a combination of classifier ensembling, synthetic data rephrasing, and reduced reliance on heuristic filters. When training 8B parameter models for 1T tokens, using a high-quality subset of our data improves MMLU by 5.6 over DCLM, demonstrating the efficacy of our methods for boosting accuracies over a relatively short token horizon. Furthermore, our full 6.3T token dataset matches DCLM on MMLU, but contains four times more unique real tokens than DCLM. This unlocks state-of-the-art training over a long token horizon: an 8B parameter model trained for 15T tokens, of which 7.2T came from our dataset, is better than the Llama 3.1 8B model: +5 on MMLU, +3.1 on ARC-Challenge, and +0.5 on average across ten diverse tasks. The dataset is available at this https URL

Subjects: Computation and Language (cs.CL)

Cite as: arXiv:2412.02595 [cs.CL]

(or arXiv:2412.02595v1 [cs.CL] for this version)

https://doi.org/10.48550/arXiv.2412.02595

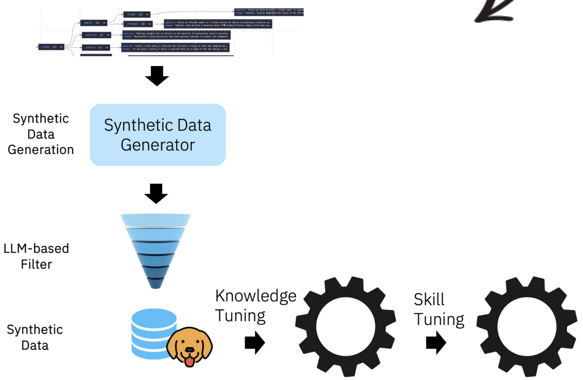
Where Synthetic Data is being used in the LLM Pipeline

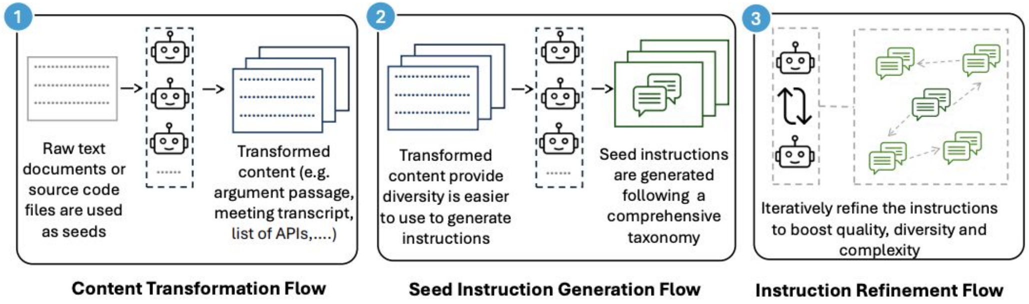
You've likely already interacted with it!

Stage of Model Development	Use Case
Pre-Training for Foundation Models	Refining, classifying, and filtering the web to remove duplicated data, biases, and in general curate better-performing datasets (as well as address the model collapse challenge).
Post-Training and Fine-Tuning	Generating instruction, preference, etc tuning datasets through seed data, in order to perform domain-specific tasks. LLMs acting as annotators can greatly speed up this process!

Synthetic representations of **post-training data**

Red Hat + IBM's InstructLab, an implementation of a student teacher approach





Raw text documents or source code files are used as seeds

Transformed content (e.g. argument passage, meeting transcript, list of APIs,....)

Transformed content provide diversity is easier to use to generate instructions

Seed instructions are generated following a comprehensive taxonomy

Iteratively refine the instructions to boost quality, diversity and complexity

Content Transformation Flow

Seed Instruction Generation Flow

Instruction Refinement Flow

Microsoft's AgentInstruct, with multiple stages of data transformation

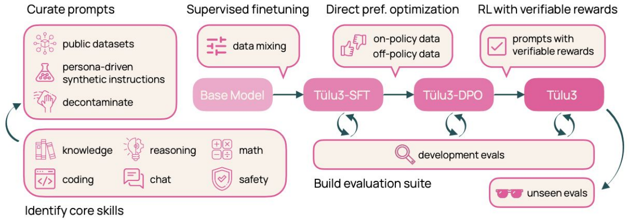


Figure 1: An overview of the TÜLU 3 recipe. This includes: data curation targeting general and target capabilities, training strategies and a standardized evaluation suite for development and final evaluation stage.

Tülu3, a multi-stage training suite

Where Synthetic Data is being used in the LLM Pipeline

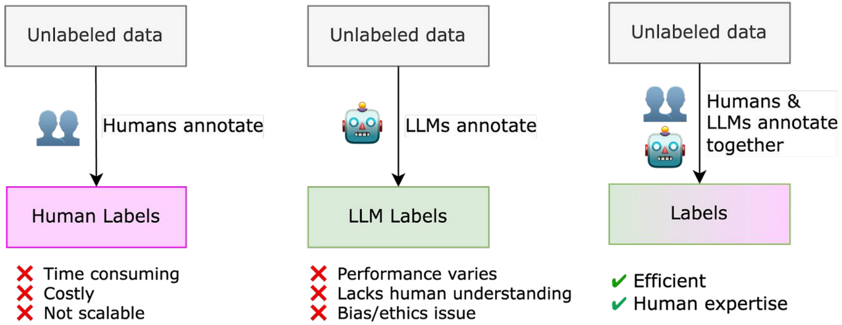
You've likely already interacted with it!

Stage of Model Development	Use Case
Pre-Training for Foundation Models	Refining, classifying, and filtering the web to remove duplicated data, biases, and in general curate better-performing datasets (as well as address the model collapse challenge).
Post-Training and Fine-Tuning	Generating instruction, preference, etc tuning datasets through seed data, in order to perform domain-specific tasks. LLMs acting as annotators can greatly speed up this process!
Model Evaluation, RAG Evaluation, etc	LLMs as judges, for example in benchmarks, but also expanding edge scenarios for RAG use cases & reducing hallucination risks.

& LLMs can greatly help generate it!

Using Generative AI for Synthetic Data

LLMs are incredibly powerful in annotating, and rapidly developing custom datasets



Unlabeled data

Humans annotate

Human Labels

- ✗ Time consuming
- ✗ Costly
- ✗ Not scalable

Unlabeled data

LLMs annotate

LLM Labels

- ✗ Performance varies
- ✗ Lacks human understanding
- ✗ Bias/ethics issue

Unlabeled data

Humans & LLMs annotate together

Labels

- ✓ Efficient
- ✓ Human expertise

Augment existing datasets to enhance the diversity and balance of data

Reduce time and resources needed for data collection and annotation

Overcome data limitations where existing data may be limited or hard to obtain

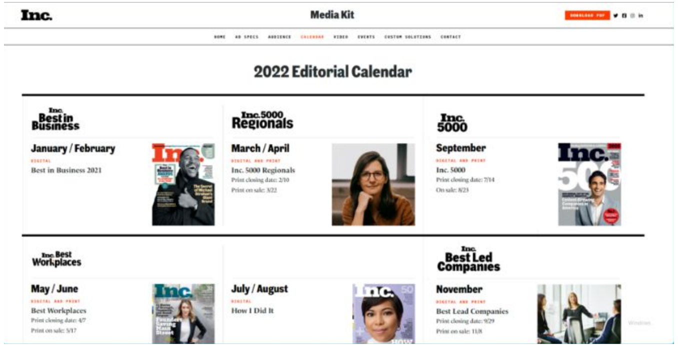
Source:
https://arxiv.org/abs/2108.13487

Enough slides, it’s time for our first demo!

Let’s learn how to customize a BERT-based model for sentiment analysis

- **1133x cheaper**, $2.7 compared to $3061
- Emits around **0.12 kg CO2** compared to very roughly 735 to 1100 kg CO2 with GPT-4
- Latency of **0.13 seconds** compared to often multiple seconds with GPT-4
- Performing **on par with GPT-4** at identifying investor sentiment (both 94% accuracy and 0.94 F1 macro)

Analyze a large news, customer, or review corpus



Inc. Media Kit

DOWNLOAD PDF

HOME AD SPECS AUDIENCE CALENDAR VIDEO EVENTS CUSTOM SOLUTIONS CONTACT

2022 Editorial Calendar

Inc. Best in Business
January / February
DIGITAL
Best in Business 2021

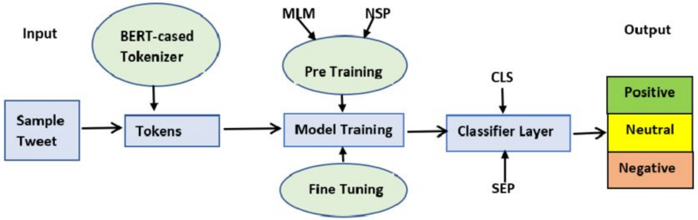
Inc. 5000 Regionals
March / April
DIGITAL AND PRINT
Inc. 5000 Regionals
Print closing date: 2/10
Print on sale: 3/22

Inc. 5000
September
DIGITAL AND PRINT
Inc. 5000
Print closing date: 7/14
On sale: 8/23

Inc. Best Workplaces
May / June
DIGITAL AND PRINT
Best Workplaces
Print closing date: 4/7
Print on sale: 5/17

July / August
DIGITAL
How I Did It

Inc. Best Led Companies
November
DIGITAL AND PRINT
Best Lead Companies
Print closing date: 9/29
Print on sale: 11/8



Source:
https://huggingface.co/blog/synthetic-data-save-costs
https://github.com/MoritzLaurer/synthetic-data-blog

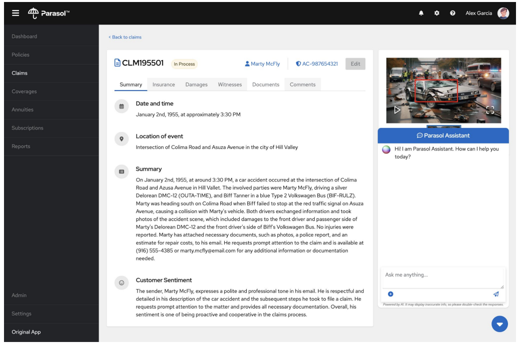
Demo #2

Let’s also learn about conversational SLM’s!

We’ve customized a domain-specific model for an insurance use case

- ➤ **~90% cheaper** versus GPT-4 blend of 80% input, 20% output tokens
- ➤ Similar reduction in CO2 usage and latency compared to generic 3rd party LLM API
- ➤ Secure and auditable **data privacy built-in**
- ➤ With InstructLab, the full tuning process doesn’t require extensive data science expertise

Assist agents in claim efficiency and reduce backlog



CLM195501 In Process

Marty McFly

AC-987654321

Date and time

January 2nd, 1955, at approximately 3:30 PM

Location of event

Intersection of Colima Road and Asuza Avenue in the city of Hill Valley

Summary

On January 2nd, 1955, at around 3:30 PM, a car accident occurred at the intersection of Colima Road and Azusa Avenue in Hill Vallet. The involved parties were Marty McFly, driving a silver Delorean DMC-12 (OUTA-TIME), and Biff Tanner in a blue Type 2 Volkswagen Bus (BIF-RULZ). Marty was heading south on Colima Road when Biff failed to stop at the red traffic signal on Asuza Avenue, causing a collision with Marty's vehicle. Both drivers exchanged information and took photos of the accident scene, which included damages to the front driver and passenger side of Marty's Delorean DMC-12 and the front driver's side of Biff's Volkswagen Bus. No injuries were reported. Marty has attached necessary documents, such as photos, a police report, and an estimate for repair costs, to his email. He requests prompt attention to the claim and is available at (916) 555-4385 or marty.mcfly@email.com for any additional information or documentation needed.

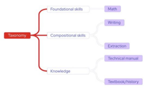
Customer Sentiment

The sender, Marty McFly, expresses a polite and professional tone in his email. He is respectful and detailed in his description of the car accident and the subsequent steps he took to file a claim. He requests prompt attention to the matter and provides all necessary documentation. Overall, his sentiment is one of being proactive and cooperative in the claims process.

Hi! I am Parasol Assistant. How can I help you today?

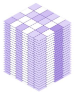
Powered by AI. It may display inaccurate info, so please double-check the responses.

InstructLab



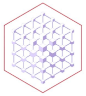
Taxonomy-driven Data Curation

Folder structure with Q&A pairs for topics to teach a model



Large-scale Synthetic Data Generation

Generate additional training data to expand dataset automatically



Model Training with New Data

Phased, large-scale alignment tuning (with knowledge and skills)

Source:
https://arxiv.org/abs/2403.01081
https://github.com/rh-rad-ai-roadshow/parasol-taxonomy

Thank you

Join the DevNation

Red Hat Developer serves the builders. The problem solvers who create careers with code. Let's keep in touch!

- Join Red Hat Developer at **developers.redhat.com/register**
- Follow us on any of our social channels
- Visit **dn.dev/upcoming** for a schedule of our upcoming events

Red Hat Developer

Build here. Go anywhere.

in linkedin.com/company/red-hat

▶ youtube.com/user/RedHatVideos

f facebook.com/redhatinc

twitter.com/RedHat